BOLT BERANEK AND NEWMAN INC

CONSULTING • DEVELOPMENT • RESEARCH

Report No. 2189
Job No. 11545

INFORMATION PROCESSING MODELS AND COMPUTER AIDS FOR HUMAN PERFORMANCE

SEMIANNUAL TECHNICAL REPORT NO. 1, SECTION 1

TASK 1: SECOND-LANGUAGE LEARNING

30 JUNE 1971

ARPA ORDER NO. 890, Amendment No. 6

Sponsored by the Advanced Research Projects Agency
Department of Defense, under Air Force Office of
Scientific Research Contract F44620-71-C-0065

Prepared for:

Air Force Office of Scientific Research
1400 Wilson Boulevard
Arlington, Virginia 22209

47

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author)
Bolt Beranek and Newman Inc.
50 Moulton Street
Cambridge, Massachusetts 02138

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP

3 REPORT TITLE
INFORMATION PROCESSING MODELS AND COMPUTER AIDS FOR HUMAN PERFORMANCE TASK 1: SECOND-LANGUAGE LEARNING

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)
Scientific Interim

5 AUTHOR(S) (First name, middle initial, last name)
Daniel N. Kalikow
Kenneth N. Stevens

6 REPORT DATE
30 June 1971

7a. TOTAL NO. OF PAGES
46

7b. NO. OF REFS
3

8a. CONTRACT OR GRANT NO F44620-71-C-0065

b. PROJECT NO 890

c. 61101D

d. 681313

9a. ORIGINATOR'S REPORT NUMBER(S)

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
AFOSR - TR - 71 - 2846

10 DISTRIBUTION STATEMENT
Approved for public release;
distribution unlimited.

11 SUPPLEMENTARY NOTES
TECH, OTHER

12. SPONSORING MILITARY ACTIVITY
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH (NL)
1400 WILSON BLVD
ARLINGTON, VIRGINIA 22209

13. ABSTRACT

Progress is reported in four areas of activity. We describe a new type of display that will allow students to receive automatic feedback on their pronunciation of minimally different pairs of words. We present the basic hardware configuration for the Mark II version of the Automated Pronunciation Instructor, and explore the implications of its expanded capabilities. We discuss our relationship with DLI, and outline our current estimates of future demands on their facilities. Finally, we report on our investigation of the pronunciation problems encountered by English speakers in learning Mandarin Chinese, in preparation for the coming field test at DLI-Monterey.

Report No. 2189 Bolt Beranek and Newman Inc.

INFORMATION PROCESSING MODELS AND COMPUTER AIDS FOR HUMAN PERFORMANCE

SEMIANNUAL TECHNICAL REPORT NO. 1, SECTION 1
Task 1: SECOND-LANGUAGE LEARNING

30 June 1971

by

Daniel N. Kalikow
Kenneth N. Stevens

ARPA Order No. 890, Amendment No. 6
Sponsored by the Advanced Research Projects Agency,
Department of Defense, under Air Force Office of
Scientific Research Contract F44620-71-C-0065

Prepared for
Air Force Office of Scientific Research
1400 Wilson Boulevard
Arlington, Virginia 22209

TABLE OF CONTENTS

SEMIANNUAL TECHNICAL REPORT NO. 1, SECTION 1
PERIOD 1 JANUARY 1971 - 30 JUNE 1971

ARPA Order No. 890
Program Code No. 1D20
Contractor: Bolt Beranek and Newman Inc.
Effective Date of Contract: 1 January 1971
Contract Expiration Date: 31 December 1971
Amount of Contract: $340,461
Contract No. F44620-71-C-0065
Principal Investigators: John A. Swets
Daniel N. Kalikow
Mario C. Grignetti
Duncan C. Miller
Telephone No. 617-491-1850
Title: Information Processing Models and
Computer Aids for Human Performance

TASK 1: SECOND-LANGUAGE LEARNING

1. Technical Problem

The task is to carry out the final development of a computer-based system for automated-instruction of the new speech sounds of second languages, and to field-test this system in the Defense Language Institute (DLI) instructional environment.

2. General Methodology

Laboratory experiments and field evaluations.

3. Technical Results

Progress is reported in four areas of activity. We describe a new type of display that will allow students to receive automatic feedback on their pronunciation of minimally different pairs of words. We present the basic hardware configuration for the Mark II version of the Automated Pronunciation Instructor, and explore the implications of its expanded capabilities. We discuss our relationship with DLI, and outline our current estimates of future demands on their facilities. Finally, we report on our investigation of the pronunciation problems encountered by English speakers in learning Mandarin Chinese, in preparation for the coming field test at DLI-Monterey.

4. Department of Defense Implications

Language schools of the Department of Defense give instruction in approximately 65 languages to over 200,000 students each year. The systems under development are designed to facilitate this instructional process.

1. PREFACE

The present contract is a partial continuation of a research program begun in 1966 under ARPA sponsorship. Of the four tasks eventually funded under Contract F44620-67-C-0033, with the Air Force Office of Scientific Research, the first two tasks were awarded continuing support under the present contract. Those tasks are:

1. Second-language learning
2. Models of man-computer interaction

The present Semiannual Technical report covers the progress made in the first of these tasks during the first six months of the new contract. We have bound the reports of the two tasks separately to facilitate their distribution and use. In addition to a copy of this page, both sections of this report contain an appropriate subset of the documentation data required for the whole report: a contract information page, a summary sheet for the particular task at hand, and a DD form 1473 for document control

2. INTRODUCTION[1]

The objective of this research is to carry out further development of a computer-based system for automated instruction in the pronunciation of a second language, and to field test this system in the Defense Language Institute instructional environment. A prototype version of this system, called the Automated Pronunciation Instructor (API) was developed and tested under AFOSR contract F44620-67-C-0033. The results of that program were described in that project's final report.

The present research program involves a three-year effort, the final two years of which will be primarily devoted to DLI field testing. The current year is to be devoted primarily to the preparation of the Mark II version of the API, and to the other scientific and administrative work that must precede the field tests. Existing software for aiding Spanish-speaking students in acquiring the speech sounds of English is to be expanded. Year 2 is to be spent primarily in field-testing the system at DLI's English Language Branch at San Antonio, for the Spanish-English language pair. Year 3 is to incluae a field-test at DLI-Monterey, where the system will be used to aid English students in learning the sounds of Mandarin Chinese.

There are four general types of activity required during Year 1. This report, covering the first six months of work on this contract, deals with them in the following order:

[1] The author gratefully acknowledges the contributions of the other members of the project team: Dennis H. Klatt, Kenneth N. Stevens, John A. Swets, and Douglas W. Dodds. He thanks Barbara A. Noel for monitoring the experiments, and Karl S. Pearsons and Sanford A. Fidell for aiding the data analysis.

A. New software development

B. Design of the Mark II API

C. Interaction with DLI

D. Investigation of the pronunciation problems encountered by English language speakers in learning Mandarin Chinese.

3. NEW SOFTWARE

Daniel N. Kalikow

At the conclusion of the experiments with the Mark I API, several areas for potential improvements in its hardware and software had become apparent. These were enumerated in the final report on the preceding project. There are preparations underway for the implementation of most of the previously suggested changes in the coming system. Several of these will be discussed in the context of Chapter 4, on new hardware; but since we have not yet completed the Mark II, this report is not the appropriate forum for a full treatment of these plans.

In the months preceding the dismantling of the Mark I API, while data analysis from the previous experiments was proceeding, we made several attempts to simulate within this limited system some of the software improvements planned for the newer system. These simulation attempts were fruitful in two senses: they demonstrated the feasibility of a basically new type of display, and they gave us valuable programming experience for this display type. The following is a description of this software as implemented on the Mark I API. Projected applications of the new approach with the Mark II will be briefly sketeched later.

It was pointed out in the final report of the previous contract that the nature of the aspiration-voice onset display (AVOD) was basically different than that used in the vowel- and reduced-vowel tongue position displays (VTPD and RVTPD). Feedback in the AVOD is presented in the form of a time graph, while the VTPD and RVTPD present their feedback in a two-dimensional space of tongue location, with time the parameter separating successive inferences about tongue position. This provides complete information about the gesture shape produced in an utterance, but at the expense of detailed feedback about the time course of the gesture. Also, due to the limitations of the type of feedback computable, both tongue-position displays are only able to produce target areas that a successful utterance must touch, but gesture shape and time course cannot be evaluated simultaneously. Because of the apparently greater strength of the AVOD in improving Ss' pronunciations, it is an obvious next step to apply this type of display in the tongue-position context.

This was a fairly trivial task in terms of what was already available within existing software in the Mark I API. When speech input is entered into the computer via the filter-bank and multiplexer-A/D converter, all the various algorithms previously produced could be applied to the contents of each spectral frame, and each algorithm produced a numerical displayable output for each time sample. In the VTPD, the tongue height and front-back outputs for each sample were used as coordinates for a display point. It is a simple step to dissociate the two axes and plot each value explicitly as a function of time.

Because of the simplicity of this task, it was implemented in a larger context than that used for the previous mode of interaction between student and machine. Recall that the single-utterance mode was exclusively employed within the Mark I API, and that this mode has certain unavoidable limitations such as inflexibility and repetitiousness. Recall further that one of the major aims for the coming system is to add the minimal-pair paradigm to the training procedures. Therefore, we investigated the feasibility of producing software that would produce *time-plotting feedback* for *minimal pairs*, for both tongue-position and aspiration functions.

3.1 EXPERIMENTAL PROGRAM

Because the pronunciation of *two* words by the student is basically incompatible with the hardware structure of the Mark I API, the only way to test the minimal-pair mode with that system was to use the free-response rather than the normal mode of the system. For the purposes of this discussion, the two modes are formally identical in terms of the computations involved, and so the difference in time course and the absence of tape-loop activity may be neglected. A further note by way of introduction: the figures to be shown in illustration of the new capability were produced from the speech of a normal English talker. Therefore, the minimal pairs used will actually result in differences in the displays. It is to be assumed that if the minimal pair is appropriate for the Spanish-English language pair (i.e., if the difference between the two words is a difficult one for a Spanish speaker to reproduce), then the differences observed below would not be apparent, and some type of feedback algorithm would be called in to inform the student of the lack of discrepancy. The present software does not contain such algorithms, but some possible procedures for their implementation will be described later.

3.1.1 Aspiration

Figure 1 was produced by the talker through utterance of the minimal pair "team - deem." The vowel involved is present in Spanish, and so would not be expected to cause any difficulty. The presence of aspiration in the first word would be a problem, and "deem" should be straightforward. As speech proceeds, output from the filter-bank is read into the machine and a display is built up on the CRT. Time proceeds to the right, with

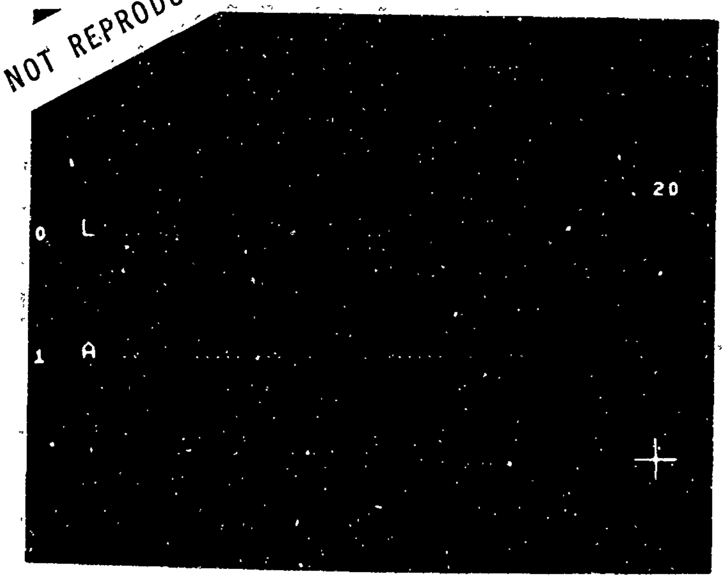


FIGURE 1. ANALYSIS OF MINIMAL PAIR "TEAM - DEEM"

(See text for explanation)

each successive point being drawn from a sample taken 20 msec later than the previous one. After each of the samples is taken, the computer begins application of its sum-and-difference algorithms to the contents of the various filters, and produces outputs interpretable in terms of various physical parameters of the vocal tract at that time. There are at least six of these algorithms in existence at present, but not all are displayed simultaneously; a *subset* relevant to the particular pronunciation problem tapped by the current minimal pair is selected for display. For the pair "team - deem," only two are required.

The top trace, labeled "L," indicates the loudness of the speech signal for each time sample during the utterance. Each sample's loudness value is given by the sum of the logarithms of the energy in the filter-bank across all channels, with the addition of certain correction terms whose purpose is to make the output proportional to what is termed "vocal effort." The corrections are inserted to compensate for the fact that utterances of the vowels /a/ (got) and /u/ (boot), with equal subjective loudnesses, have different physical energies. A simple energy indicator would, therefore, discriminate between vowels, and this is undesirable for the intended purpose. In the present displays, loudness is always plotted at the top, to give the student an indication as to the time course of the speech sample. It may be used here to determine that at first there was silence, then a word, then another period of silence, then another word, then finally silence.

The second trace, labeled "A," indicates aspiration intensity. This is computed as the simple sum of activity in filters 14 through 17, without most of the temporal constraints used in the AVOD and described in the previous report. If an aspiration point is above baseline, this indicates that there is activity in the relevant filters *without* the low-frequency activity associated with voicing in vowels. That is, aspiration is displayed only before voice-onset time at the beginning of words and after voice offset time at their conclusion. This latter condition accounts for the apparent aspiration activity at the conclusion of both spoken words—during the /m/—and will certainly be eliminated in future versions of this program by a simple logical addition that was not incorporated in this program for reasons of time. To emphasize the presence of aspiration activity, the baseline is not plotted for samples producing non-zero aspiration values.

Let us concentrate our attention on the beginnings of each word. It is clear that there is considerable aspiration activity associated with the first word, and none with the start of the second. If this minimal pair were incorporated into a full-scale training regimen for aspiration of initial stops in Spanish speakers, the job of the students would be to maximize the difference in aspiration between the two words. A probable starting condition for a speaker with severe problems would be no aspiration in either word, or, if present, for roughly similar amounts to be demonstrated for both words. It would be a simple matter to devise a logical rating procedure which would compute the ratio or difference between the aspirations displayed for the pair, and to reinforce the student accordingly. Alternatively, the analysis of a correct minimal pair as produced by a teacher could be presented to the student for his inspection and comparison. Both of these avenues will be explored in the Mark II API.

Since the sampling rate was 50/sec (one each 20 msec) and there are a total of 50 points plotted for each algorithm applied to the speech of this minimal pair, a one-second period of speech has been analyzed and displayed. We realize that this sampling rate is too slow to follow transients in the speech—10 msec, as in our previous work, is correct—but since memory space in the Mark I was limited, we elected to simulate additional time-storage capability by stretching the sample rate. Whereas with all previous work the same 50 spectral frames referred to only one-half second of speech, in minimal-pair mode we could investigate speech periods twice that length at the cost of some sacrifice in detail. This will be unnessary in the coming system, since memory will be greatly expanded.

A further example of the response of the system to minimal pairs sensitive to aspiration problems is given in Fig. 2. Here, the English speaker uttered the word pair "coat - goat" with the system configured identically to the above description. The Loudness profiles of the two words are quite similar. Note especially the separation between the /o/ and the /t/, caused by the momentary stop preparatory to the aspirated /t/ concluding the words. The Aspiration profiles of the two words are very dissimilar, and can be made more so if the points plotted by the final /t/'s are disregarded (recall that this can be easily done by inserting a condition on Aspiration that it only be displaved *before* voicing onset in a given word). Again, there is considerably more aspiration observed at the start of the first word than at the start of the second, and feedback algorithms such as mentioned above may be easily added.

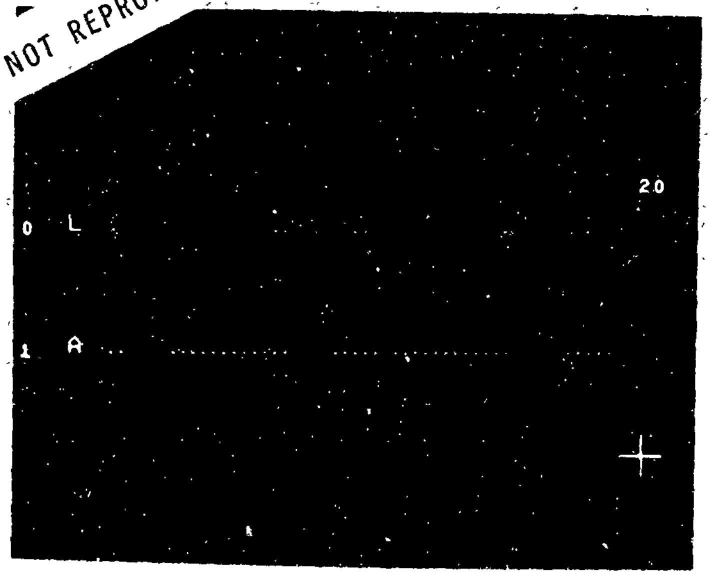


FIGURE 2. ANALYSIS OF MINIMAL PAIR "COAT - GOAT"

3.1.2 Vowels

The above displays are new primarily in the sense that they have added minimal pairs to the previously described AVOD algorithm. The aspiration algorithm was used in time-plotting mode, and the time course of the two-word gesture was displayed. However, tongue-position information has never been presented in this manner. Figure 3 shows the analysis of the word-pair "boat - bought," as spoken by the same English talker. The Loudness trace is shown, as before, to orient the viewer to the positions of the words on the time trace. The terminal /t/'s are obvious as triangular bursts of points at the conclusions of the words. The primary humps of the words arise from the vowel portions, and are indistinguishable in the Loudness algorithm since it is sensitive to overall energy regardless of spectral content. The difference between the two vowels, /o/ and /ɔ/, is made clear by the trace below, labeled "H - L." This is the display of the high-low tongue-position algorithm, proportional to the value of the first formant.[1] The difference between the two vowels in tongue height is clear in this speech sample. Presumably, a Spanish speaker having difficulty with this vowel pair would be able to produce an acceptable rendition of "boat," since /o/ is present in Spanish; but his rendition of /ɔ/ would be too similar to /o/ in tongue height. Evaluation and reinforcement algorithms, and or comparisons with analysis of teacher utterances, can be implemented along the lines described above for the aspiration display.

[1] It is computed for each time sample, but is displayed only for those time samples whose filter 2 energy is above a threshold value. This constraint ensures the display of tongue-position points onlyfor those samples during which a vowel is present, as indicated by voicing. If a displayed point is above the line, it indicates tongue height above its resting position; and vice-versa.

NOT REPRODUCIBLE

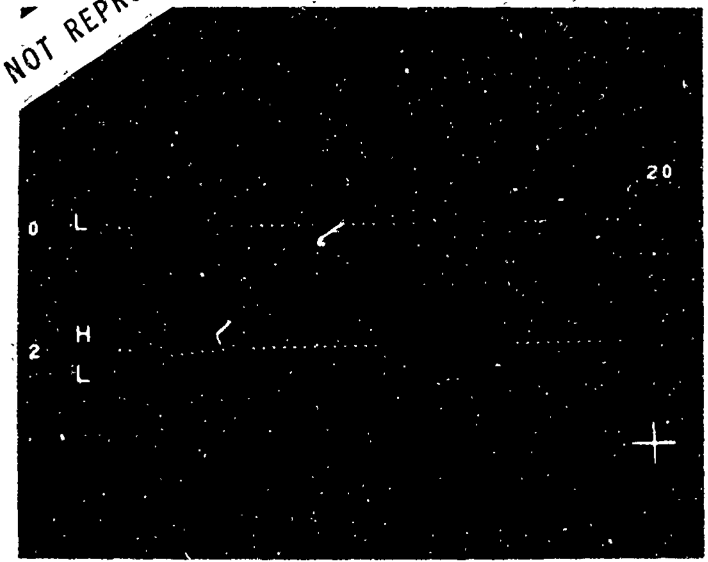
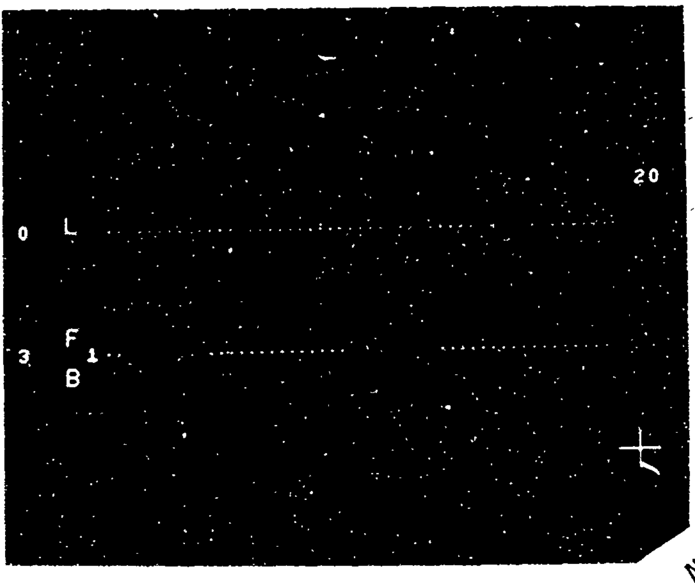


FIGURE 3. ANALYSIS OF MINIMAL PAIR "BOAT - BOUGHT"

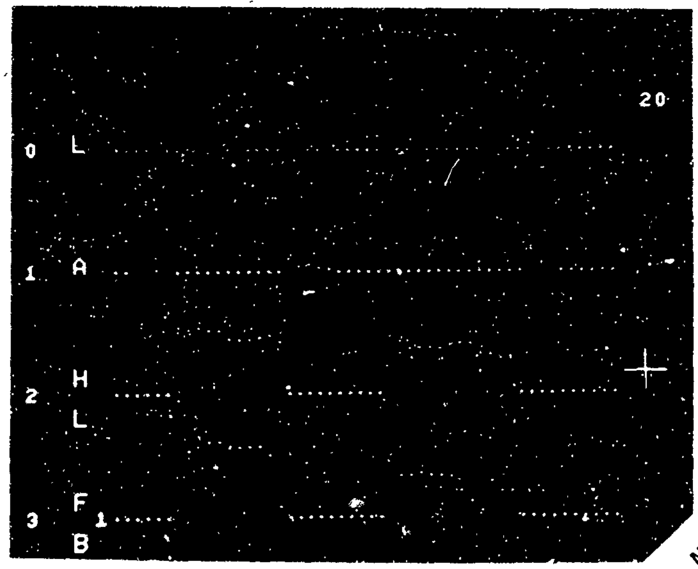
Figure 3 showed the analysis of a word pair whose two vowels differed primarily in terms of tongue height. The front-back discrepancy is of less importance for that word pair, and so it presumably would not be shown to S to avoid distracting him. Figure 4 presents the opposite side of the coin: the analysis of a successful utterance of the pair "beet - bit." The Loudness profiles again show the terminal /t/'s, and no difference between the vowels in terms of effort profiles. Since the primary difference between the /i/ and the /I/ vowels is along the front-back dimension, only the F - B function is plotted below. The difference in "frontness" of the two vowels is apparent, and amenable to the same reinforcement paradigms outlined above.

To illustrate the great capabilities for real-time speech analysis made possible in the Mark I, we have included Fig. 5. It is, admittedly, too complex for display to naive Ss, but it may help to make clear some of the operations described above, since they are all subsets of the same operating system. Here, the analysis of the word pair "team - dim" is given, with all above systems operating simultaneously. The Loudness trace indicates successful aspiration at the beginning of the first word, but not in the second; this is correct. At the same time, the viewer receives information about the tongue-positions inferred for the vowel portions of the utterances. As in Fig. 3, the vowel pair is /i/ - /I/; and, confirming the notion that the primary difference between these vowels is tongue "frontness," we find a greater difference in the F-B plot for the two vowels.



NOT REPRODUCIBLE

FIGURE 4. ANALYSIS OF MINIMAL PAIR "BEET - BIT."



NOT REPRODUCIBLE

FIGURE 5. ANALYSIS OF MINIMAL PAIR "TEAM - DIM."

Undoubtedly, the implementation of this approach in the Mark II API will differ considerably in terms of the outward appearance of the display and in the behavior of the student in calling out and evaluating it. The basic computational procedures will nevertheless remain, perhaps overlaid with simplifying constraints on the detail presented to S, and certainly with a faster speech sampling rate made possible by the expanded memory capabilities of the newer processor. Additionaly, the larger memory will make possible the analysis of longer segments of speech, by both student and recorded teacher, by hardware improvements to be further described below.

New normalization procedures are now being considered, as arc methods for adding feedback on the proper durations of voiced portions of the words. This latter type of feedback will probably be similar to that previously used in the AVOD.

4. NEW HARDWARE

Daniel N. Kalikow

Hardware and software are inextricably interrelated in any computer-based system, and especially so in the present context. Therefore, a more complete idea of the coming displays may be obtained through a consideration of the hardware configuration upon which they will be based.

At this writing, the Mark I API has been disassembled. Its central processing unit has been reassigned to other experimental duties at BBN, and its various peripherals are in various stages of reworking for incorporation within the Mark II, or they have been discarded. The Mark II itself is well on the way to completion, at least as far as the home system is concerned. There remain certain design decisions on the exact configuration of the field system which can only be made after inspection of the operation of the first-completed API, and the field system to be evaluated at DLI schools is at an earlier stage of construction.

The basic plan of the two Mark II API's is identical, and so they will be discussed as one. The design decisions mentioned above are with regard to the exact configurations of some of the custom-built peripherals whose basic operations are independent of their physical layouts.

Figure 6 is a block diagram of the Mark II API arranged to indicate the input-output structure of the system. Figure 7 is topologically identical, but it is more realistic in that it displays the components in their approximate physical relationships. Actual operation of the system is most easily explained

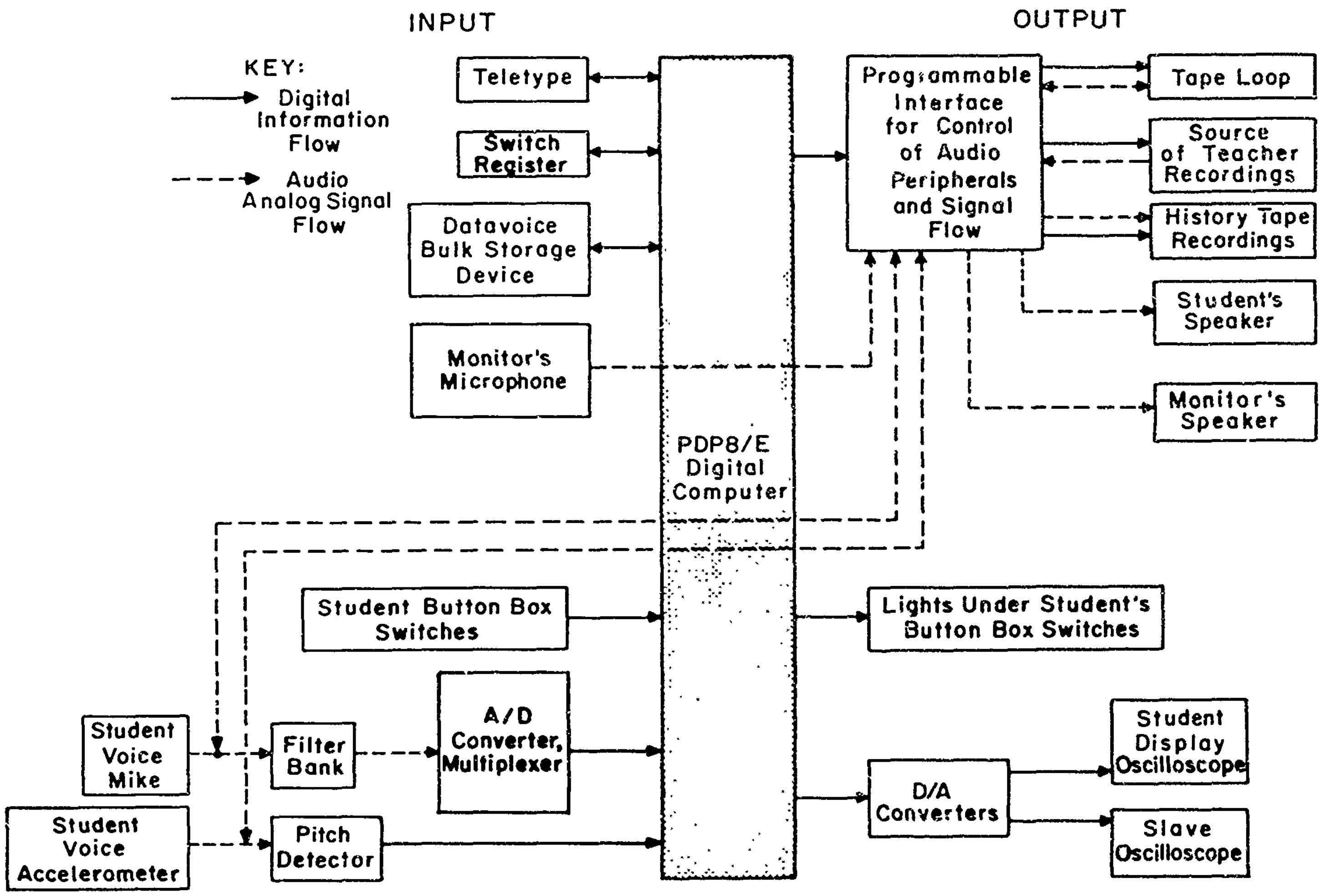

INPUT
OUTPUT
KEY:
Digital Information Flow
Audio Analog Signal Flow
Teletype
Switch Register
Datavoice Bulk Storage Device
Monitor's Microphone
PDP8/E Digital Computer
Progɿammable Interface for Control of Audio Peripherals and Signal Flow
Tape Loop
Source of Teacher Recordings
History Tape Recordings
Student's Speaker
Monitor's Speaker
Student Button Box Switches
Lights Under Student's Button Box Switches
A/D Converter, Multiplexer
Filter Bank
Pitch Detector
Student Voice Mike
Student Voice Accelerometer
D/A Converters
Student Display Oscilloscope
Slave Oscilloscope

FIGURE 6. BLOCK DIAGRAM OF BBN AUTOMATED PRONUNCIATION INSTRUCTOR, MARK II: INPUT-OUTPUT STRUCTURE

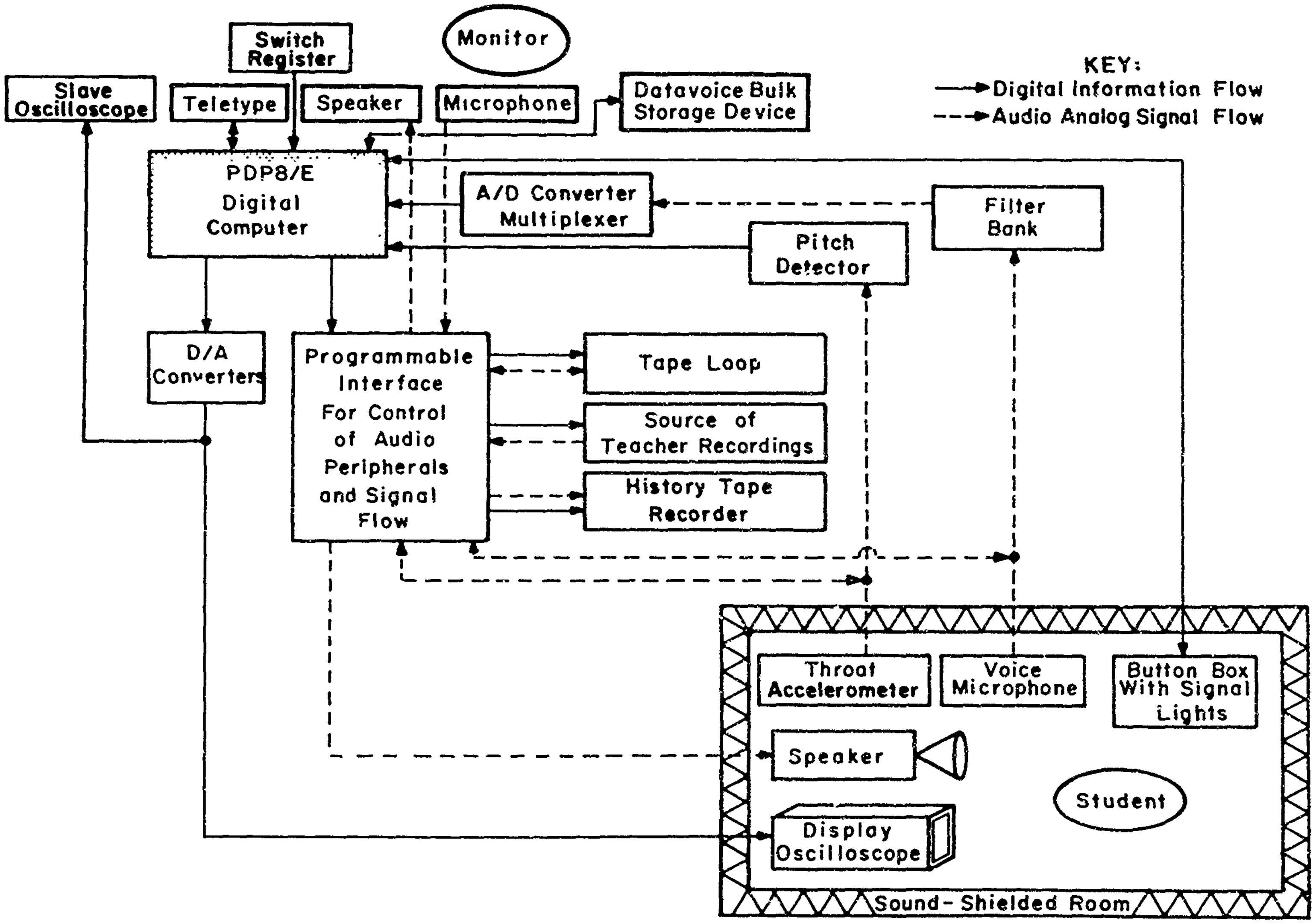


FIGURE 7. BLOCK DIAGRAM OF BBN AUTOMATED PRONUNCIATION INSTRUCTOR, MARK II: PHYSICAL ARRANGEMENT OF EQUIPMENT

in terms of Fig. 7, and the components will be introduced and described in roughly counterclockwise crder, as far as is possible. Appropriate references to the physical arrangements specified in Fig. 7 will be made. Keep in mind that all components not contained in the sound-shielded student room are mounted in the main equipment racks of the API, adjacent to the processor.

The central unit in the API is, of course, the PDP-8/E digital computer. It has a direct role in every digital signal path, and exercises direct control over audio signals through its external interfaces.

There are four information channels through which the monitor (M) interacts with the API as he controls the system. He uses the teletype to record student identification on the summary records, to enter any necessary paper-tape information relevant to that student, and to interact as needed with the executive program during training. The switch register, an integral part of the computer, is used to control operation of the Datavoice, to initiate various programs, and to select constant modes of operation within a given program while it is running. The Datavoice is a high-capacity bulk storage device. It is basically a tape-recorder modified to handle digital information in a form directly compatible with the core configuration of the computer, and it can store a complete 8K program in about 15 feet of tape, just a few seconds of playback time. It is used to speed up the availability of many separate training programs within the same API unit.

The final input of the Monitor is a microphone, with which he can converse with S during a session without entering S's room. If S has a question about procedure or if some verbal interaction is required for any reason, either S or M may initiate the exchange. The signal passes through amplification and digitally controlled interfacing, to emerge at S's speaker. Similarly, signals from S's voice microphone are passed through to M's speaker to complete the loop.

The student enters information into the system in three major ways. The first of these illustrated in Fig. 6 is the button-box. This is one of the major design improvements in the Mark II. New features include integral mounting of the buttons flush with the surface of a student's chair-desk combination; better human-engineering of the button configuration, with color-coding and the possibility for alphameric labels adjacent to each button; and finally, internal programmable lights within each button, with which the executive program may instruct S as to which buttons he may operate next. This will considerably lighten S's cognitive load during training and will allow increased concentration on the speech analysis.

There are two major paths through which S's speech passes into the API for analysis and recording. A high-quality head-mounted microphone picks up the standard acoustic signal from the vicinity of S's mouth. The signal is routed into the audio circuitry for recording, and into the filter-bank for acoustic analysis. The recording process itself will be described below;

the acoustic analvsis performed by the filter-bank is identical to that employed in the Mark I. The conversion of the spectral samoles into stored logarithmic energy values within the computer will be more efficient in the Mark II.

The extension of the API approach to the English-Mandarin Chinese language pair necessitates the provision of feedback concerning the fundamental frequency of the voice. Such information is not available from the filter-bank outout, since its purpose of deriving the spectral envelope is at direct variance with the extraction of fundamental frequency from the speech signal. It is known that available approaches to the extraction of the fundamental from conventional audio recordings of speech are quite expensive, and also that they are not totally reliable. The other commonly employed strategy for extraction of the fundamental, or pitch extraction, involves the placement of a second microphone at the throat of the speaker, to pick up the glottal activitv determining the pitch at a point upstream of the considerable modifications imposed by the rest of the vocal tract. We are currently developing a new type of pitch detector using an extremely lightweight accelerometer which can simply be taped to the throat of the student, eliminating the need for the uncomfortable elastic straps required for standard throat microphones. This transducer provides a clear glottal signal quite easily converted into digital terms for production of instantaneous voice pitch feedback. This system is still under development and will be reported on more fully in the future. Suffice it to say that the voice pitch information derived from the throat transucer travels in two paths similar to those traversed by the voice signal: it is passed into analog recording equipment for possible playback and/or later analysis, and it is directly en-entered into the system via the link between the pitch detector logic and the external bus of the computer.

On the output side of Fig. 6, we first encounter the CRT displays produced by the API for the instruction of the student. S's display scope is the sole medium through which the acoustic analyses of his speech are presented to him. It is larger than the slave display scope mounted in front of M, since M needs it only to follow the progress of the experiment while S needs to devote detailed attention to it. The large display scope being used is the Hewlett-Packard 1300A. It is fast, silent, and has an excellent contrast ratio over its ample rectangular screen.

We have mentioned the roles of S's and M's speakers in conversational interactions. Of course, the *main* role of S's speaker is to provide audio feedback for him in the manner similar to that used in the Mark I. S listens to the contents of the tape loop, comparing his utterances to those of the teacher. Teacher utterances are previously recorded and available during training as reference stimuli.

Aside from the new computer and the addition of the pitch detector, the greatest improvement in the system hardware is in the area of the tape loop and teacher recordings. It will be recalled that the Mark I required S to cycle through a stack of Language Master cards containing teacher utterances of the training materials. This tended to cause errors within the session, as S sometimes lost his place or mis-entered a card. The fidelity of the Language Master unit was also marginal. This made it impractical to attempt any acoustic analysis of teacher speech for use as a template by S. The tape loop used was a modified Ampex tape drive whose fidelity was unavoidably degraded by the mod\ification. It also was without any means of directly sensing tape position. This latter limitation made tape loop control unnecessarily difficult.

All of the above problems have been substantially reduced by our acquisition of tape transport units custom-fabricated from standard modules, by the Mackenzie Laboratories of El Monte, California. Each API will have one of these units, composed of two endless-loop cartridge transports. Each transport controls the movement of a two-track tape; and the playing of any unit is automatically terminated when the equipment senses the passage of a reflective foil over a photocell. One of the two transports will take on the duties of the tape loop. One section of the loop will be used by S as he repeatedly records the training materials; the other section will contain a teacher recording which he will hear following each utterance in store mode. The second transport will take on the role of the Language Master unit in the Mark I. It will contain as many foil-delimited sections as there are different training utterances. Each section will contain two tracks of information about a teacher's utterances: a voice track for audio feedback and possible concomitant acoustic analysis of spectral structure, and a pitch track for on-line analysis of the pitch structure of the utterances. When S wants to move to the next training stimulus, all he need do is press the appropriate button and the next available section of teacher utterance will be dubbed over onto the teacher's half of the tape loop. When both units automatically stop, the tape loop will be positioned for the first student recording of the material onto the loop. When this is done, the loop unit is switched automatically to playback, and S hears the teacher recording of the utterance he has just attempted. Switching paths for this playback information to be passed through the filter bank and pitch-detector logic will be provided, making it possible to analyze the acoustic characteristics of the teacher recordings in ways outlined above.

As before, both models of the Mark II have history tape recorders, used as a repository of the vocal behavior of S during each session. The major difference in the history tapes will be that two tracks of information will be available: one for voice output, the other for throat, or voice-pitch output.

The programmable interface interposed between computer and all audio equipments serves the purpose of allowing software control of the interconnections between the various units. This is a custom-made unit whose heart is a bank of 12 single-pole, double-throw switches whose positions may be changed by execution of the appropriate instruction by the computer. It is to be used for connecting microphones to tape recorders, tape playbacks to speakers, and other similar tasks. The home system API interface has been fabricated in a plug-in format, to retain design flexibility while the proper interconnection scheme is worked out in practice. Following this design phase, the field system interface will be hard-wired to ensure reliability and promote economy of construction.

It is quite possible to replicate all existing routines that were operational on the Mark I API within the new hardware configuration of the Mark II. However, this would be an incomplete utilization of its powers. An example of the kind of software made possible by the hardware can be found within a consideration of the minimal pair mode. A possible tape loop appropriate for that mode might have four subsections delimited by the photoelectric cue marks: two for S's pair, and two for the teacher's versions of the word. Following utterance of each word, the transport might stop to give S a breather before the next. The increased speech storage time and concurrent analysis of teacher's utterances would produce a display logically similar to the time-course plots presented above by the Mark I, but with considerably more imposed schematization.

Since this software has not yet been developed, it is not appropriate at this time to speculate further along these lines. It is, nevertheless, hoped that the preceding presentation of the design capabilities of the hardware, and of some outlines of the planned software, has given the reader some sense of the new teaching procedures that will take shape during the coming months.

5. COORDINATION WITH DLI

Daniel N. Kalikow

The goal of this development program is, of course, the production of a field system of the API (and of a software package) suitable for evaluation of the effectiveness of the API approach within the DLI context. To this end, more is required than a physical machine and the programs to make it produce useable feedback: we must begin active cooperation with DLI personnel to prepare for the introduction and evaluation of the system at DLI-San Antonio and DLI-Monterey.

A first step toward this goal of cooperative development was taken in May 1971. Under the auspices of the Air Force Office of Scientific Research, a meeting was held to inaugurate the interaction between Bolt Beranek and Newman Inc. and the Defense Language Institute. Representing BBN were Drs. Kalikow and Swets of this project; Drs. Levin and Allardyce represented DLI. Dr. Hutchinson observed for ARPA-AFOSR, and he had also invited Drs. Hyman, Burgess, and Miss House from BESRL, since their research is related to the present project.

From our point of view, the purpose of the meeting was to orient the group to the aims of the project, and to begin the elicitation of the new information about the DLI system and environment that would be essential to the mission's completion. We perceived the need for a period of general orientation toward the present and projected capabilities of the API approach, since we feel that such an understanding would foster realistic expectations on DLI's part of the nature and scope of the research effort that could be mounted. Having concluded this, we turned our collective attention toward the consideration of those practical administrative matters that could profitably be considered at that stage.

As stated at the meeting, there are three areas in which DLI cooperation is essential. These areas are summarized below in the form mutually agreed upon at the meeting. This is not to say that the substantive requests are not subject to revision or that they have been granted; rather, what follows is the basis for a formal request to the appropriate DLI channels for assistance, subject to negotiation and change. Still, these requests have passed through at least one iteration of the BBN-DLI interaction, and so we have cause to believe that the plans expressed below can be implemented practically.

5.1 STUDENT SAMPLE

This is probably the crucial need of the project, since, of course, one of the prime reasons for moving toward the DLI system was its possession of a controlled teaching environment with students of tested capabilities. We needed confirmation that the numbers of students we contemplated running, and the projected training times, would accord with the daily schedules of students and with the arrival times of new batches of students at the Institute.

Our projected need is based on the following basic premises. It is the plan of this project to hire and train a field-technician for the API system. This person will not be continuously available to monitor the equipment, and that places an upper limit on the amount of time per day that the system can be operative. A more stringent limit on available time is placed by the tight schedule of the DLI students. It is, of course, necessary to minimize interference with the standard language-teaching program set up by DLI. Because of this, it may be difficult to find more than four class hours in each day from which DLI administration considers it possible to excuse students. Giving each student a full hour at the machine at any one time, and staggering two groups of four students on alternating class days for a total of 12 sessions per student, or 24 consecutive class days, yields a total number of eight students passed through the program in five weeks' time. Successive replications of the procedure on successive classes, or within the same class of a long-term DLI course, would multiply the number of students testing the system. Assuming that a control treatment not using the API system is devised and used, all available training time would be used in running actual experimental students. There would be no limit on the numbers of control students used, save that of the evaluation statistics themselves.

The above student capacities are, of course, hypothetical. In terms of orders of magnitude, though, they are fairly accurate, since the one-station nature of the present API system and the scheduling complexities of the DLI class structure place unchanging limits on the training procedures. Parenthetically: it may prove advisable to encourage students to use the system during their free hours, in a less structured manner than the strict usage schedule outlined above. This could make it possible to run a few more students than the above projections. Having considered all these projections for numbers of students needed, Drs. Levin and Allardyce expressed the opinion that DLI would have no difficulty in providing sufficient numbers of students for any evaluation program of the above order of magnitude, and encouraged our formal submission of requests to that end, following such further consultations as might be appropriate.

5.2 TECHNICAL CONSULTATION

We requested, and received, information concerning personnel within the DLI system with whom we might discuss problem areas in pronunciation and teaching of pronunciation for the two subject language pairs.

We began, and will continue, discussion of issues and procedures in student selection, specification of proper control groups, and instruments for evaluation of the effectiveness of training: in short, consultation on experimental design was begun.

We requested assurances of the availability of assistance, if needed, in the preparation and translation of instruction booklets and other written training materials, since this may be a problem in our preparation of the Spanish-English software. We were told that this is possible given a reasonable forewarning and amount of work. Similarly, we were told that trained Chinese Mandarin informants would be made available for production of appropriate recordings of training materials for that program, and that assistance in curriculum preparation could also be obtained, as above. In summary, then, DLI was and will continue to be responsive to the needs of the project in this area, and we expect to follow up on this expressed willingness within the next few months.

5.3 PERSONNEL AND FACILITIES

Since the present project is a complex undertaking, it will be of value to have a DLI liaison man available at both DLISA and DLIWC, with whom BBN and our field technician may interact. This request was deemed reasonable and appropriate by the DLI representative, and will be acted upon at the appropriate time.

At both sites, office space for the field technician will be necessary, as will access to secretarial facilities for help in scheduling the students, and in maintaining the flow of correspondence and data between the site and BBN's home laboratory, where data will be treated and where backup software will be prepared.

The following physical space requirements for the API system itself were outlined: two adjacent rooms, at least one of which shall be quiet and sound-shielded to prevent interference with student speech analysis. It shall be possible to run cables between the rooms, preferably through a wall. A one-way glass window between rooms would be helpful to the monitor. Appropriate 120 volt power and air conditioning of both rooms is necessary.

As was their reaction to the preceding areas of needed BNI-DLI cooperation, the reaction of the DLI representatives to this last category of requested assistance was straightforward and positive. It was a good beginning to our association, a meeting that augured well for the further consultations that will be undertaken in the coming months.

6. A PRELIMINARY STUDY OF SOUNDS OF CHINESE-MANDARIN

Kenneth N. Stevens

We report here on the current status of our investigation of the primary pronunciation problems to be expected in English-speaking students learning Mandarin Chinese. This work is still in progress, and so what follows is subject to expansion and revision as more data become available.

Probably the most difficult aspect of Mandarin Chinese pronunciation for a native speaker of English is the production of the tones. In the part of this project concerned with Mandarin Chinese, therefore, attention will be focused primarily on the development of displays and accompanying training procedures to facilitate learning of the tones.* At present, there are only meager quantitative data on the contour shapes and frequencies for the Mandarin tones (Wang and Li, 1967; Howie, 1969; Chuang, *et al.*, 1971), although there is general agreement among phoneticians as to the description of the tones. As a first step in the formulation of training procedures based on visual display, therefore, it is appropriate to obtain measurements of the tones and to assemble some information concerning the range of variation that is permissible in various aspects of the fundamental frequency (F_o) contours for each tone, if the tone is to be acceptable to a native speaker. Since there is evidence that at least some tones are influenced by the phonetic context in which they occur, particularly by the context of other tones, it is

*We expect also to examine certain sounds and contrasts in Mandarin Chinese that are not familiar to speakers of English. These include the contrasting voiceless fricatives (syǎu, śau, sháu) and affricates (chǎu, chyǎu and jyàu, jàu), and the zero finals (e.g., dž, sz̀, and jŕ, chr̄). We have not yet established whether these (and possibly other) contrasts in Mandarin Chinese present sufficient difficulties to English learners of Mandarin to warrant the development of displays for these contrasts.

necessary to collect data on the tones not only as they occur in isolated syllables but also in combination with other syllables.

With these objectives in mind, we have begun a study of the shapes of the fundamental-frequency contours for the Mandarin tones as produced by several native speakers. This study includes the collection of recordings of items in an appropriately selected list of syllables and phrases in Mandarin, making of narrow-band spectrograms of each of these utterances, and analysis of these spectrograms to obtain the fundamental frequency versus time, together with data on the durations of the vowel segments and the timing of other articulatory events. At the present time, we have analyzed data from only one speaker and have recorded a list of utterances for several additional speakers.

6.1 GENERAL INFORMATION ON THE TONES OF MANDARIN CHINESE

The phonological structure of Mandarin Chinese is based on the syllable. It is convenient to regard every syllable as composed of an initial, which may have one or more consonants or no consonant, and a final, which consists of a vowel, a sequence of vowels, or a sequence consisting of vowels and sonorant (nasal) consonants. The final part of the syllable may have one of four tones. These tones are described qualitatively in the following manner:

Tone 1. This tone stays level on a relatively high pitch, and is written ā.

Tone 2. A syllable with this tone starts in the middle register and rises. It is designated by á.

Tone 3. A syllable with this tone starts at a relatively low pitch. It stays relatively level for an interval of time, or may drop slightly initially, and this low-pitched phase is then followed by a rise. This tone is written ǎ.

Tone 4. This tone falls sharply from fairly high to quite low, and is designated by the symbol à.

6.2 PRELIMINARY MEASUREMENTS OF THE TONES IN ISOLATED SYLLABLES

Typical contours for each of the four tones as they occur in isolated syllables are shown in Fig. 8. A logarithmic frequency scale (slightly distorted at low frequencies below about 70 Hz) is used in this and subsequent displays. The general form of these contours is consistent with the phonetic descriptions. Tone 1 is level and at a high frequency—about 180 Hz for this speaker. Tone 2 ends at about this frequency, and the onset of tone 4 also occurs at this frequency or slightly higher. Tone 3 remains almost level in the range 80-90 Hz before rising abruptly near the end of the syllable.

Averaging durations of the syllables, from time of release of the initial consonant to end of voicing are given in Table 1. Syllables with tone 3 are much longer than the others, those with tones 1 or 2 are of intermediate duration, and tone-4 syllables are the shortest. These relative durations are consistent with the data of Chuang, *et al.* (1971).

TABLE I. Average durations of monosyllables in Mandarin Durations are from release of initial consonant (or from onset of voicing when there is no initial consonant) to termination of voicing. Averages are for 4-5 utterances, one speaker.

	msec
Tone 1	360
Tone 2	370
Tone 3	600
Tone 4	240

Tones 2 and 3 both begin at a relatively low frequency and then rise. The contours indicate that tone 3 starts at a lower frequency than tone 2, and F_0 tends to fall before the final rise. Syllables (in isolation) with tone 3 are also much longer than those with tone 2. This difference in duration is entirely in the initial low-frequency part of the tone. Measurements show that, on the average, the time taken for the contour to increase in frequency by 10 percent relative to the starting frequency by 10 percent relative to the starting frequency is 150 msec for tone 2 and 480 msec for tone 3.

6.3 CONTOURS WHEN THE TONES ARE FOLLOWED BY SYLLABLES WITH ZERO TONE

When an utterance has several syllables, some syllables may have no tone at all, i.e., carrying "zero" tone. A syllable of this kind is usually joined to the immediately preceding syllable, and its fundamental frequency is, in some sense, a continuation of that of the preceding tone. Examples of each of the four tones followed by a zero-tone syllable are shown in Fig. 9. In

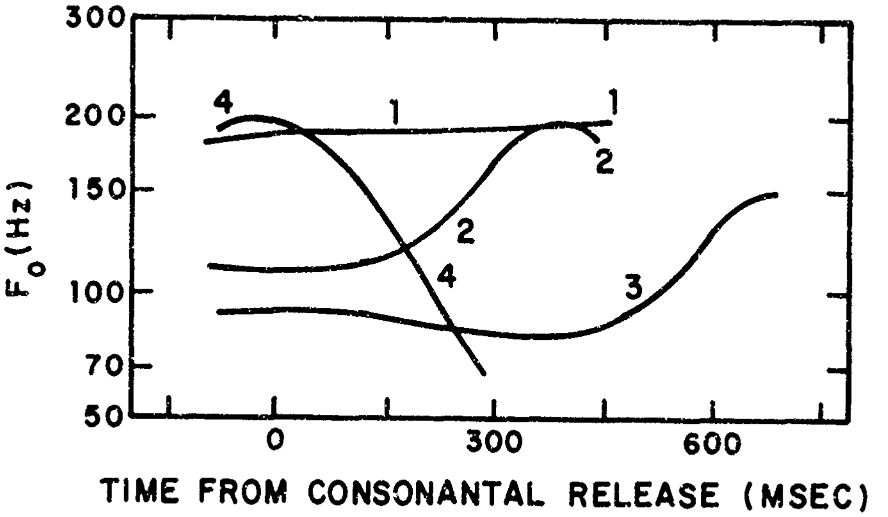


FIGURE 8. TYPICAL CONTOURS OF FUNDAMENTAL FREQUENCY (F_o) *VERSUS* TIME FOR SINGLE SYLLABLES ILLUSTRATING THE FOUR TONES OF MANDARIN CHINESE. THE UTTERANCES ARE: (1) lā, (2) lá, (3) lǎ, (4) là.

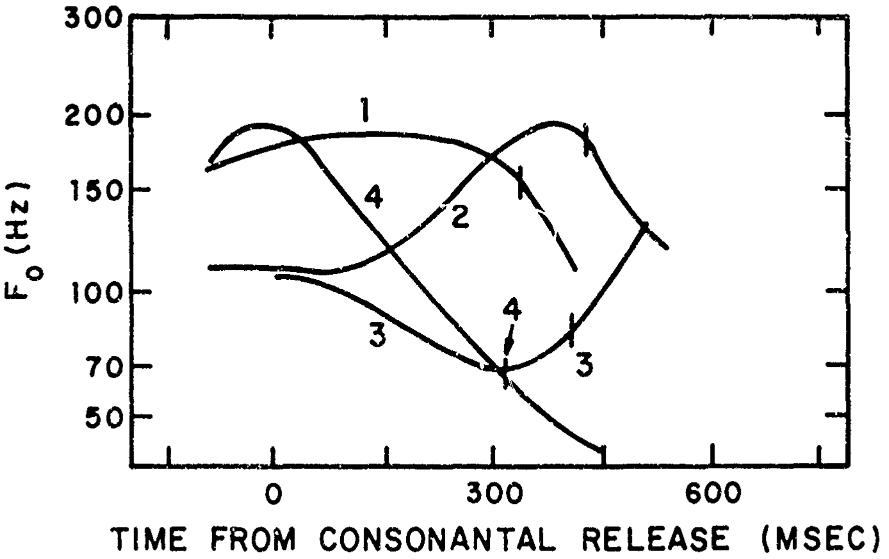


FIGURE 9. TYPICAL CONTOURS OF F_o *VERSUS* TIME FOR TWO-SYLLABLE UTTERANCES IN WHICH THE SECOND SYLLABLE CARRIES ZERO TONE. THE UTTERANCES ARE: (1) lāle, (2) lále, (3) fěile, (4) làle. THE SHORT VERTICAL MARKS IDENTIFY THE RELEASE OF THE FINAL [l]

the case of tones 1, 2 and 4, the fundamental frequency for the final zero-tone syllable shows a decrease relative to the frequency at the termination of the preceding syllable. For tone 3, however, the zero-tone syllable has a fundamental frequency that is higher than that at the end of the preceding syllable. In fact, the contour for the entire two-syllable combination beginning with tone 3 is similar to that for the tone 3 in isolation, the final syllable carrying the rising portion of the contour.

6.4 SEQUENCES OF TWO SYLLABLES EACH CARRYING A TONE

Some contours of F_0 versus time for typical sequences of two syllables are shown in Fig. 10. If for the moment we exclude from consideration the sequence of tone 3 followed by tone 3, then several general comments can be made about contours of this type.

1. The durations of the syllables in initial position are much shorter than those in final position. The durations of all tones in initial position are about the same (except for tone 3 followed by tone 3, as noted later).

2. The F_0 contours for the various tones in the final position for a two-syllable sequence are basically the same as for the tones in isolated syllables. The durations are somewhat shorter, however, and, at least for this speaker, the duration for tone 4 in this context is not appreciably shorter than that for tones 1 and 2.

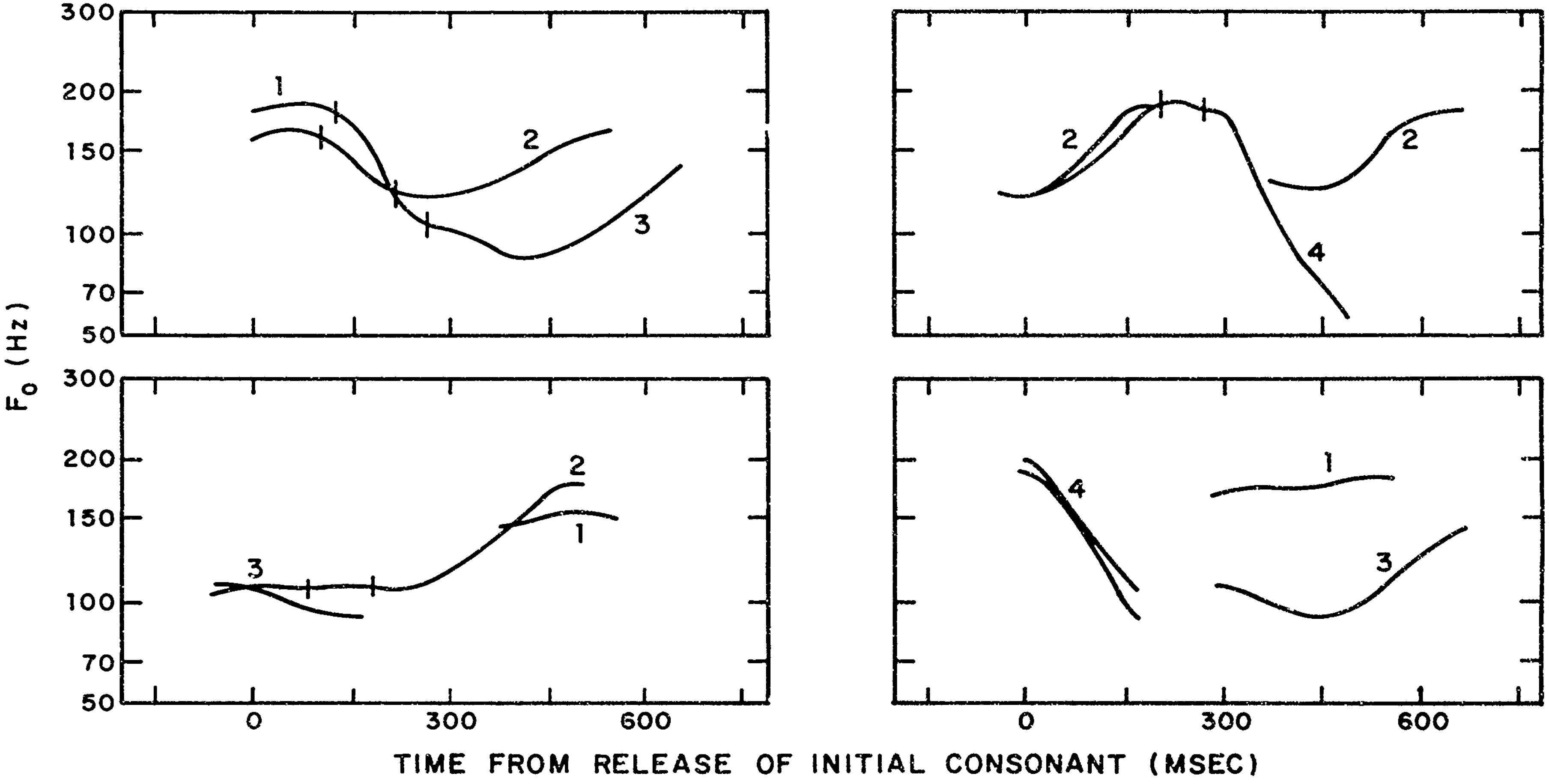


FIGURE 10. CONTOURS OF F_o *VERSUS* TIME FOR SAME TWO SYLLABLE SEQUENCES. THE UTTERANCES ARE (1-2) tā mái, (1-3) tā mǎi, (2-2) lái chá, (2-4) dū lì, (3-1) ní kāi, (3-2) bǐ níu (4-1) tài qāu, (4-3) tài hǎu. THE SHORT VERTICAL MARKS IDENTIFY CONSONANT BOUNDARIES.

3. Tone 3 in initial position can be interpreted simply as a level, low tone, without the terminal rise that is characteristic of the isolated or final syllables. Other tones in initial positions often have maximum or minimum F_0 values that are not as extreme as those of the tones in isolation.

Figure 11 shows F_0 contours for two syllable-pairs: tone 3 followed by tone 3, and tone 2 followed by tone 3. For the first of these sequences, the initial tone 3 begins with a somewhat higher F_0 than the normal tone 3, and has some of the attributes of tone 2. However, at least for this speaker, the initial tone 2 is shorter than the initial tone 3, and the two sequences can be distinguished on this basis. This observation is not entirely in accord with the findings of others, who state that the 2-3 and 3-3 sequences are similar and often indistinguishable. Further measurements with other speakers are needed to resolve this apparent inconsistency.

6.5 SOME COMMENTS ON A DISPLAY OF FUNDAMENTAL FREQUENCY AS AN AID TO TEACHING THE PRONUNCIATION OF MANDARIN TONES

Although further data are needed before the contours for the various tones in different contexts can be specified, together with the allowed variation in various aspects of the tones, a few remarks can be made concerning displays to be used as aids for teaching pronunciation of the tones. The initial plan is to use a display that shows a curve, or series of points, of F_0, on a logarithmic scale, versus time. A logarithmic scale seems appropriate since, on the basis of preliminary analysis of the contours for several speakers, it appears that the appropriate form of the contours on a logarithmic scale remains the same

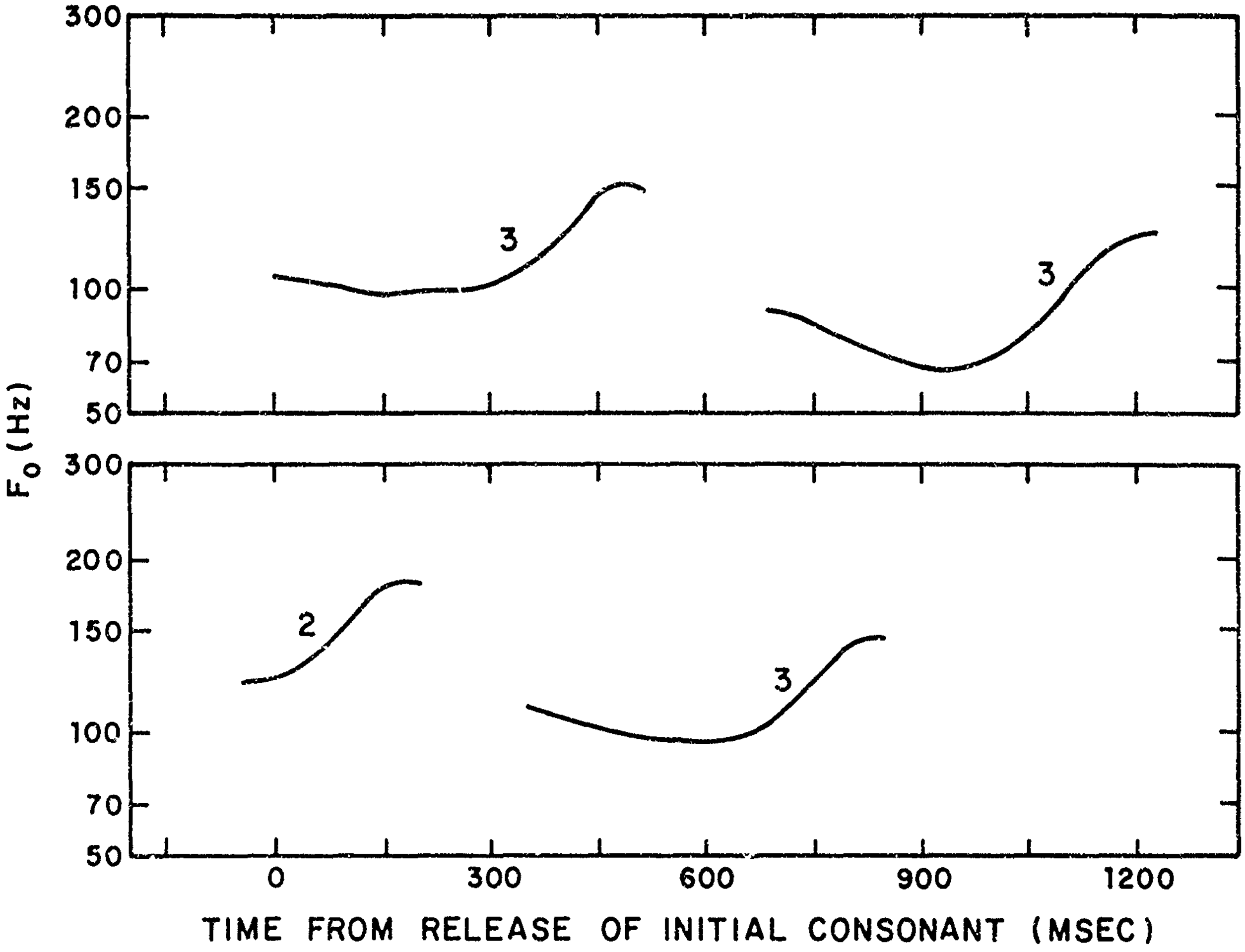


FIGURE 11. CONTOURS OF F_o *VERSUS* TIME FOR TWO-SYLLABLE SEQUENCES CONTRASTING THE TONES 2-3 (ná byǎu) AND 3-3 ([illegible]ŏ hǎu)

for speakers with different average values of fundamental frequency. That is, it appears that contours appropriate for speakers with different average fundamental frequency can be obtained simply by shifting certain prototype contours up or down on a logarithmic frequency scale.

The fact that different tones have different durations (at least in isolation and in final position) indicates that it is important to preserve not only the form of a contour but also its duration. Furthermore, since the contour for a particular tone may be influenced by adjacent tones, any training procedure should incorporate displays of F_0 for sequences of two or more tones. Teachers of Mandarin are, of course, well aware that drills involving sequences of tones, as well as syllables in isolation, are a necessary part of learning the pronunciation of this language.

7. REFERENCES

Chuang, C-K., Hiki, S., Sone, T. and Nimura, T. The acoustical features and perceptual cues of the four tones of the standard colloquial Chinese. Preprint of paper to be presented at the Seventh International Congress on Acoustics, Budapest, August 1971.

Howie, J. M. The vowels and tones of Mandarin Chinese: Acoustical measurements and experiments, Ph.D. Thesis, Indiana University, October 1970.

Wang, W. S-Y, and Li, P-P. Tone 3 in Pekinese. J. Speech and Hearing Res., 1967, 10, 629-636.